

Title : will be set by the publisher
Editors : will be set by the publisher
EAS Publications Series, Vol. ?, 2015

AUTOMATED ECLIPSING BINARY DETECTION: APPLYING THE GAIA CU7 PIPELINE TO HIPPARCOS

Berry Holl¹, Nami Mowlavi¹, Isabelle Lecoœur-Taïbi¹, Fabio Barblan¹,
 Lorenzo Rimoldini¹, Laurent Eyer¹, Maria Süveges¹, Leanne Guy¹, Diego
 Ordoñez-Blanco¹, Idoia Ruiz¹ and Krzysztof Nienartowicz¹

Abstract. We demonstrate the eclipsing binary detection performance of the Gaia variability analysis and processing pipeline using Hipparcos data. The automated pipeline classifies 1 067 (0.9%) of the 118 204 Hipparcos sources as eclipsing binary candidates. The detection rate amounts to 89% (732 sources) in a subset of 819 visually confirmed eclipsing binaries, with the period correctly identified for 80% of them, and double or half periods obtained in 6% of the cases.

1 Introduction

The Gaia mission is expected to observe ~ 1 billion sources, among which 0.4 to 7 million are expected to be eclipsing binaries (as summarized in Holl *et al.*, 2013). The detection and characterisation of those eclipsing binaries are distributed over two Coordination Units (CUs) of the Gaia Data Processing and Analysis Consortium. CU7 has the task to identify eclipsing binaries, find their orbital periods, and sub-classify them. This information is then passed to CU4, who will model the eclipsing binaries and derive stellar and orbital parameters (Siopis & Sadowski, 2012). The processing loop is schematized in Fig. 1 of Holl *et al.* (2013).

In this short contribution, we demonstrate the eclipsing binary detection performance of the current Gaia CU7 pipeline on the Hipparcos data (ESA, 1997). Because the Hipparcos time sampling and mean number of observations are similar to Gaia, it is a good dataset to test the performance of the Gaia processing pipeline.

The Hipparcos data set is described in Sect. 2 and the results are presented in Sect. 3. Conclusions are drawn in Sect. 4.

2 Hipparcos eclipsing binaries

We define a reference set of 819 Hipparcos eclipsing binaries after visual inspection of the light curves, chosen among the list of eclipsing binaries published in Vol. 11 of the “Hipparcos catalogue of periodic variable stars” of ESA (1997), to which

¹ Department of Astronomy, University of Geneva, CH 1290 Versoix (e-mail: berry.holl@unige.ch)

| Module | All Hipparcos | | Eclipsing binaries ref. set | |
|-----------------------------|---------------|---------|-----------------------------|--------|
| | #sources | % | #sources | % |
| Input time series | 115 423 | 100.00% | 819 | 100.0% |
| Variability detection | 15 568 | 13.49% | 819 | 100.0% |
| Supervised clas.: eclipsing | 1 598 | 1.38% | 752 | 91.8% |
| SOS Eclipsing binaries | 1 067 | 0.92% | 732 | 89.4% |

Table 1. Pipeline processing result from top to bottom for the whole Hipparcos catalog (left), and for the eclipsing binaries reference set (right).

additional Hipparcos sources are added that are flagged as eclipsing binaries in the January 2014 revision of the AAVSO catalogue (Watson *et al.*, 2013). We must note that the periods published in the Hipparcos catalog were derived from Hipparcos light curves for only 682 sources. The periods published for the remaining 137 sources were taken from the literature. We thus do not expect our automated pipeline to recover the periods of all those later sources.

3 Gaia CU7 pipeline applied to Hipparcos

The Gaia CU7 pipeline, outlined in Eyer *et al.* (2015), can be divided in four per-source analysis steps. They are briefly described in the following sections, with our application to the 115 423 Hipparcos sources that have at least one good observation (flag 0 or 1) in their time series. Note that the applied pipeline configuration is simplified with respect to what is planned for official (Gaia) data processing.

3.1 Variability detection

Variable sources are detected using a χ^2 test with a p-value threshold of $p < 10^{-4}$. This reduces the list of time series to be processed to 15 568, see Table 1.

3.2 Characterisation

Periodic sources are searched using the unweighted Lomb-Scargle method (Lomb, 1976; Scargle, 1982), and multi-harmonic Fourier series are fitted to their light curves.

3.3 Classification

The 15,568 variable sources are classified into 23 different variable types using a supervised Random Forest classifier (Breiman, 2001). Input to the classifier are various Hipparcos specific attributes (detailed in Dubath *et al.*, 2011) which are derived from the light curve model parameters determined in the previous step, together with the parallax, and V-I colour. Eclipsing binaries are represented by one class containing types ‘EA’, ‘EB’, and ‘EW’. We base our training set on Dubath *et al.* (2011), which however includes 72% of the eclipsing binaries in our reference data set (Sect. 2). To make the reference set more independent of the training set, we train our classifier with only half of the Dubath training set, containing only 37% of the reference eclipsing binaries. Although this reduces the precision of the classifier, it strengthens the power of the reference set to evaluate the pipeline performance. The (ten-fold cross-validation) confusion matrix of the

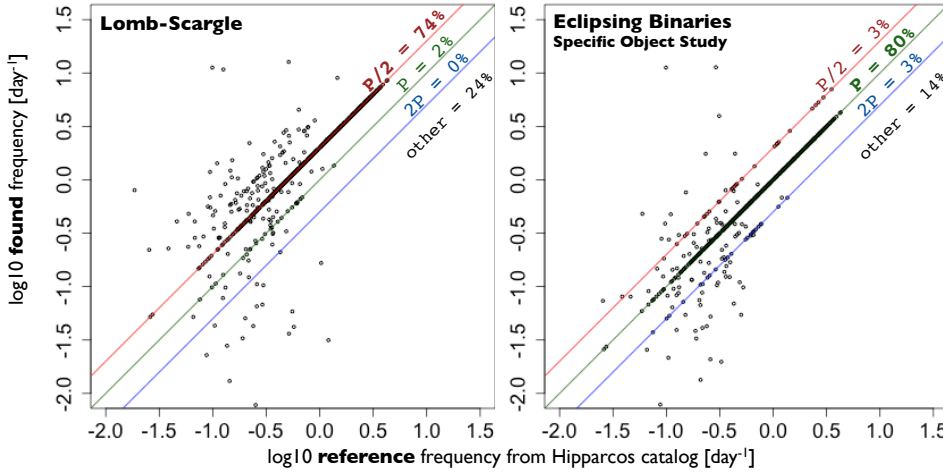


Fig. 1. Period recovery for the 732 eclipsing binaries from the reference set that were identified as eclipsing binaries. The results from the unweighted Lomb-Scargle period search (left), and the ‘corrected’ periods using our Specific Object Study module ‘Eclipsing Binaries’ (right) recovering the correct period in most cases.

training set has a high completeness¹ of 94.7% and a contamination² of only 7.5% for eclipsing binaries.

Table 1 shows that, applied to all Hipparcos data, the trained Random Forest classifier predicts 1 598 sources to be of type eclipsing binary (selecting those with probability > 0.5), which includes 92% of the reference eclipsing binaries.

3.4 Specific Object Studies (SOS): Eclipsing Binaries

This post-classification step processes all 1 598 eclipsing binary candidates provided by supervised classification. It aims at finding the best period and at sub-classifying the eclipsing binaries based on a two-Gaussian model description of the eclipses in the folded light curves. The detection and period identification algorithms will be detailed in Holl *et al.* (in prep.), and the two-Gaussian modeling and sub-classification algorithm in Mowlavi *et al.* (in prep.). Basically, the automated procedures test the goodness of the model fits for several fractions and multiples of the computed Lomb Scargle period. If not satisfactory, the procedure is repeated for additional periods found with phase-dispersion minimisation (Jurkevich, 1971; Stellingwerf, 1978; Schwarzenberg-Czerny, 1997) and String Length (Lafler *et al.*, 1965; Burke, 1970). The best period is retained from those tests. The eclipsing binary is then sub-classified based on the two-Gaussian model parameters as described in Mowlavi *et al.* (in prep.).

This last step of per-source pipeline processing confirms 1 067 eclipsing binaries (Table 1), of which 89% (732) are in our reference set of eclipsing binaries. Fig-

¹Type completeness = number of correctly classified sources / number in training set.

²Type contamination = 1 - number of correctly classified sources / number classified.

ure 1 shows that of these 732 reference eclipsing binaries, SOS Eclipsing Binaries recovers³ the periods listed in the Hipparcos catalog for 80% of the sources, and double or half periods for 6%. As mentioned in Sect. 2, the periods in the Hipparcos catalog were taken from literature for 137 sources. The SOS Eclipsing Binaries package identifies 113 of those and recovers the period for 48%, and double or half the period for 7%.

4 Conclusions and discussion

Applying the Gaia CU7 pipeline to the 118 204 Hipparcos time series, 1 067 (0.9%) are identified as eclipsing binary candidates. Validating the results against a reference set of 819 visually identified eclipsing binaries we find that 732 (89%) are included, for which the Hipparcos period is recovered in 80% of the cases. Assuming that the reference set contains all eclipsing binaries detectable in Hipparcos data translates into a completeness of 89% and contamination of 31% of our 1 067 candidates. Further investigation of this 31% ‘contamination’ is planned to be included in Holl *et al.* (in prep.).

Given the similarities between Hipparcos and Gaia observations, the good detection rate on Hipparcos data suggests that the current CU7 variability analysis and processing pipeline is in good shape to automatically detect eclipsing binaries in Gaia too.

References

- Burke *et al.*, 1970, J. Roy. Astron. Soc. Canada 64, 353
 Breiman L., 2001, Mach. Learn., 45, 5
 Dubath, P., Rimoldini, L., Süveges, M., *et al.* 2011, MNRAS, 414, 2602
 Dworetsky, M.M., 1983, MNRAS 203, 917
 ESA, 1997, The Hipparcos and Tycho Catalogues, ESA SP-1200
 Eyer, L., Evans, D. W., Mowlavi, N. *et al.* these proceedings
 Holl, B., *et al.*, in preparation
 Holl, B., Mowlavi, N., Eyer, L., *et al.* 2013, EAS Publications Series, 64, 399
 Jurkevich, I., 1971, Astrophys. Space Sci. 13, 154
 Lafler, J., Kinman, T.D., 1965, ApJ Suppl. 11, 216
 Lomb, N. R. 1976, Astrophys. Space Sci, 39, 447
 Mowlavi, N., *et al.*, in preparation
 Scargle, J. D. 1982, Ap. J., 263, 835
 Schwarzenberg-Czerny, A. 1997, ApJ, 489, 941
 Siopis, C., & Sadowski, G. 2012, Orbital Couples: Pas de Deux in the Solar System and the Milky Way, 59
 Stellingwerf, R.F., 1978, Ap. J. 224, 953
 Watson, C., Henden, A. A., & Price, A. 2013, VizieR Online Data Catalog, 1, 2027

³The period $X \cdot P_{true}$ is recovered if $|P_{true} - P_{found}/X| (\Delta T/P_{true}) < 0.1 P_{true}$, with ΔT the time-series duration and P_{true} the Hipparcos or literature period, see Dubath *et al.* (2011).